A structure-based nomenclature for Bacillus thuringiensis and other bacteria-derived pesticidal proteins

Article (Published Version)

Journal of Invertebrate Pathology xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Journal of Invertebrate Pathology

journal homepage: www.elsevier.com/locate/jip

A structure-based nomenclature for *Bacillus thuringiensis* and other bacteria-derived pesticidal proteins

Neil Crickmore[a,*], Colin Berry[b], Suresh Panneerselvam[c], Ruchir Mishra[c], Thomas R. Connor[b], Bryony C. Bonning[c]

[a] School of Life Sciences, University of Sussex, Brighton BN1 9QG, UK
[b] School of Biosciences, Cardiff University, Cardiff CF10 3AX, UK
[c] Department of Entomology and Nematology, University of Florida, Gainesville, FL 32611, USA

ARTICLE INFO

Keywords:
Cry toxin
Nomenclature
Toxin classification

ABSTRACT

In 1998 a nomenclature for the growing list of pesticidal proteins from *Bacillus thuringiensis* (Bt) was derived based solely on protein sequence comparisons. This nomenclature was widely adopted and provided a robust framework for the naming and classification of the proteins. The success of these proteins in integrated pest management schemes prompted an increased effort to find others with improved or more diverse activities. These discovery activities led to the characterization of proteins from a wider range of bacteria and with a variety of different protein folds. Since most of these new proteins were grouped together as Cry proteins it became apparent that the existing nomenclature had limitations in representing the diverse range of proteins that had been identified. This revised nomenclature retains the basic principles of the 1998 version but provides specific mnemonics to represent different structural groups. For the purposes of consistency, the vast majority of the proteins have either retained their name or have a new name that clearly references the previous one. Other pesticidal proteins not previously included in the nomenclature have been incorporated into this version.

1. Introduction

The first cloned gene encoding a *Bacillus thuringiensis* crystal protein was reported in 1981 (Schnepf and Whiteley, 1981) and, as further genes were cloned over the following years, a nomenclature for the encoded proteins was proposed (Hofte and Whiteley, 1989). In this nomenclature, proteins were classified according to their insecticidal activities, with CryI proteins being toxic to lepidopteran insects, CryIIs to both Lepidoptera and Diptera, CryIIIs to Coleoptera and CryIVs to just Diptera. Although this nomenclature proved extremely useful in systematically classifying proteins that had been previously been given arbitrary names, it soon became apparent that there were significant limitations. One such limitation was that proteins that shared sequence homology often had different insecticidal specificities, requiring them to be put into different primary classification groups. Another major limitation was the need to obtain comprehensive bioassay data before a protein could be classified. To overcome these challenges, a revised nomenclature was introduced in 1998, which classified the proteins solely by amino acid similarity (Crickmore et al., 1998). In this system, proteins were compared in a multiple sequence alignment and a dendrogram produced to illustrate their relatedness. Names were derived based on the location of the node at which the protein joined the dendrogram. A four-level naming system was adopted in which proteins that shared at least 45% sequence identity were placed in the same primary classification group (Cry1, Cry2 etc). The primary groups were then further split such that proteins that shared less than 78% identity were allocated different secondary ranks (Cry1A, Cry1B etc). A third level was used for proteins within the secondary rank that shared less than 95% sequence identity (Cry1Aa, Cry1Ab etc). Finally a fourth level was used for proteins within the same tertiary level that shared greater than 95% identity (Cry1Aa1, Cry1Aa2 etc). Although it was realised that this naming approach was potentially unstable, as more proteins were added it proved to be robust and is still used more than 20 years later. In 1998 it was recognised that there were different types of crystal protein and this led to two mnemonics being adopted, Cyt for the dipteran active proteins with a generalized *in vitro* cytolytic activity and Cry for the other crystal derived insecticidal proteins. A third mnemonic was also introduced (Vip) for insecticidal proteins that Bt secreted during vegetative growth (Estruch et al., 1996) and a further secreted toxin (SIP) was also described (Donovan et al., 2006). Within the nomenclature, it was recognised that there were a number of proteins that showed very little sequence similarity but were nonetheless

* Corresponding author.
E-mail address: n.crickmore@sussex.ac.uk (N. Crickmore).

https://doi.org/10.1016/j.jip.2020.107438
Received 7 May 2020; Accepted 5 July 2020

allocated Cry names (Cry6, Cry15 and Cry22). As more of these "outliers" were characterized they were identified as being members of specific groups (e.g. the Bin-like Toxin_10 pfam group and ETX/Mtx2-like) despite all sharing the Cry mnemonic (de Maagd et al., 2003). With the proliferation of genome sequencing projects, and improved procedures for protein structure determination, it has recently become clear that there is a wide variety of bacteria-derived insecticidal proteins and that the existing nomenclature heavily constrains the appreciation of their diversity. For this reason, the need for a classification system that better reflects structural differences has gained momentum.

2. Scope of the revised nomenclature

In the development of this more structure-based classification system, a widespread consultation exercise was undertaken involving academics, industry scientists and regulators. The following summarize the outcomes of those deliberations.

2.1. Toxins or pesticidal proteins?

Historically, the insecticidal proteins produced by *B. thuringiensis* have been referred to as Cry toxins, Bt toxins etc. As their use in biocontrol products, and in genetically modified crops, has increased, it has been observed that outside of the academic context, the use of the word toxin has negative connotations. In an attempt to mitigate this negative perception, we suggest that the preferred term should be pesticidal protein.

2.2. What counts as a pest?

Although the Cry proteins are best known for their insecticidal activity, their activity against other invertebrate targets is well established and there has been significant progress in their use against nematode pests (Hu et al., 2018). Some Cry proteins – the so-called parasporins – have activity against human cancer cell lines (Ohba et al., 2009). As there is a significant body of published research on these proteins they will be retained in the nomenclature. There is no absolute definition of what constitutes a pest when it comes to deciding whether or not a particular protein should be included. Although it is anticipated that the nomenclature will concentrate on invertebrate targets, new activities will be considered on a case-by-case basis.

2.3. Source of the pesticidal proteins

As well as limiting the range of organisms targeted by the pesticidal proteins within the nomenclature, the range of proteins included will also initially be limited by their source. The intention is to focus on proteins of bacterial origin. This would, therefore, exclude examples from sources such as spider venom (Windley et al., 2012) or plants (Liu et al., 2019). Exceptions to this may be considered on a case-by-case basis, for example if a protein with relevant activity, and clearly related to a family of proteins within the nomenclature, is characterized from a non-bacterial source.

2.4. Structure as the primary unit of classification

Having considered the scope of the project, the basis of a structure-based classification system had to be defined. As described above, proteins currently in the nomenclature already represent a number of distinct structural classes (3-domain, Toxin_10/Bin-like, ETX/Mtx2-like etc). Furthermore it is well established that proteins sharing sequence homology are likely to exhibit similar structural configurations. Thus the use of homology comparators – such as the pfam database (El-Gebali et al., 2019) can provide a reliable method of grouping sequences by predicted structure.

2.5. Maintaining relationships with the existing nomenclature

As mentioned above, the existing nomenclature has been widely accepted and adopted over the last two decades and, therefore, it is prudent not to jeopardize that bank of embedded information, or totally abandon an established vocabulary describing important compounds. The current nomenclature consists of a three letter mnemonic followed by four classification levels indicated by a mix of alphanumeric characters. To maintain the relationship between the old and new nomenclatures, the 4-level classifier for a given protein will be retained even though the mnemonic may have changed. Thus a protein (hypothetically) previously called Cry88Fa3 would become Nnn88Fa3, where Nnn represents a particular structural class. If and when the nomenclature runs out of single character symbols for the secondary classifier (e.g. Cry32Za) the primary level will be divided i.e. Cry32.1Aa will follow Cry32Za. Various options exist should the tertiary level run out of characters, including the use of Greek characters.

2.6. Will the four-level classifiers be backfilled or duplicated?

In order to avoid confusion, numbers removed from the original Cry class (e.g. Cry6, Cry15, Cry22) will never be reused as Cry proteins. The same principle will apply for Vip1, Vip2 and Vip4 which will not be reallocated. For those new classes of protein that have been derived from the Cry or Vip classes (e.g. Tpp, Mpp, Vpa – see Table 1) numbering will start at the next available number that has not previously been applied to a protein in that class, or the parent class. For other classes which have historically been distinct (e.g. Vip3 and Cyt), or have been added as distinct classes (e.g. Spp and Pra) the next available number for each class will be used. At the point of transition to the new nomenclature the highest primary rank classifier was Cry80Aa, this protein became Tpp80Aa and so new proteins within the Cry, Mpp, Tpp, Gpp, App and Xpp classes will each have started with the 81 primary classification.

3. The naming process

In deriving an efficient, robust and meaningful naming system, many approaches were evaluated. Eventually, following much testing, a very simple process was chosen. The principles and detail of this system are described below.

3.1. Sixteen structural classes initially defined

Table 1 and Fig. 1 show the initial 16 classes that have been defined. Three of these (Cry, Cyt and Vip3) are unchanged from the previous Bt toxin nomenclature. For Cyt and Vip3 all proteins previously included in these classes retain their existing names. The Cry class now only includes those proteins believed to possess the classic 3-domain structure. This includes proteins that have an extended C-terminus and those that do not, and also includes variants that contain additional regions e.g. beta-trefoil domains. The other classes represent non-3-domain proteins previously bearing the Cry mnemonic and/or new sequences that have been added (from bacteria other than Bt). These other classes are based on the listed pfam domains, known structures and other available information. The three-letter mnemonics were chosen to reflect either what type of pesticidal protein they represent (Mpp – Mtx2-like; Tpp – Toxin_10-like etc.) or some historical designation (Mcf, Mtx). One class – Xpp – has been designated as a holding class and will include proteins for which insufficient information is available to allocate them to a specific class. The Xpp mnemonic should be considered temporary.

Table 1
Classification groups within the revised pesticidal protein nomenclature. Conserved protein domains associated with each class are given along with examples and their protein database codes where known.

Class	Previous classification	Conserved domain(s)	Description (PDB codes)
Cry	Cry	pfam03945, pfam00555, cd04085	Proteins originally isolated from *B. thuringiensis* **cry**stals in which the active form normally consists of three domains. Examples include Cry1Aa (1CIY) and Cry3Aa (1DLC)
Cyt	Cyt	pfam01338	**Cyt**olytic, normally single domain, proteins such as Cyt2Aa (1CBY)
Vip	Vip3	pfam12495, pfam02018	Multi-domain proteins originally identified as being **V**egetative **I**nsecticidal **P**roteins such as Vip3Bc (6V1V)
Tpp	Cry, Bin	pfam05431	Beta pore-forming **p**esticidal **p**roteins containing the **T**oxin_10 (Bin-like) domain. Examples include Tpp35Aa (previously Cry35Aa 4JP0) and Tpp1Aa (previously BinA 5FOY)
Mpp	Cry, Mtx2, Sip	pfam03318	Beta pore-forming **p**esticidal **p**roteins from the ETX/**M**tx2 family. Examples include Mpp51Aa (previously Cry51Aa 4PKM) and Mpp2Aa (previously Mtx2)
Gpp	Cry	pfam06355	Aeg**e**rolysin like **p**esticidal **p**roteins such as Gpp34Aa (previously Cry34 4JOX)
App	Cry, Pax, Xax, Yax		Predominantly **a**lpha helical **p**esticidal **p**roteins such as App6Aa (previously Cry6Aa 5KUD) and App1Ca (previously YaxA 6EK7)
Spp		pfam01289, pfam17440	**S**phaericolysin like **p**esticidal **p**roteins
Mcf		pfam12920	Proteins related to the "**M**akes **C**aterpillars **F**loppy" toxins originally described from *Photorhabdus*.
Mtx	Mtx1		Proteins related to the **Mtx**1 toxin (2VSE) originally isolated from *Lysinibacillus sphaericus*
Vpa	Vip2	cd00233	Proteins related to the ADP-ribosyltransferase **a**ctive component of binary toxins such as **Vip**2 (1QS2) (from the Vip1 / Vip2 toxin)
Vpb	Vip1, Vip4	pfam07691, pfam03495, pfam17475, pfam17476	Proteins related to the **b**inding component of binary toxins such as **Vip**1 (6SMS), Vip4.
Pra	PirA		Proteins related to the ***P**hotorhabdus* Insect-**R**elated toxin **A** component.
Prb	PirB	pfam03945	Proteins related to the ***P**hotorhabdus* Insect-**R**elated toxin **B** component.
Mpf	PluMACPF GNIP	pfam01823	Pesticidal proteins that are part of the **M**embrane Attack Complex / **Perf**orin superfamily.
Xpp			A holding class for pesticidal proteins with currently uncharacterized structures.

3.2. Use of existing pesticidal proteins to form a scaffold for the revised nomenclature

Many different approaches were attempted to find a nomenclature system that can be easily automated while also being intuitive and meaningful. Most of these approaches produced a result that was around 90% identical to the existing classification. In terms of which method produced the best result, it soon became clear that there was no right result and that any particular method was as good or as bad as any other. With no clear biological justification for any particular method, it was concluded that the existing nomenclature structure (within a structural class) could be maintained and provide a robust scaffold for the incorporation of new sequences. When reanalysing the sequences, one did stand out as having been inappropriately assigned and so the

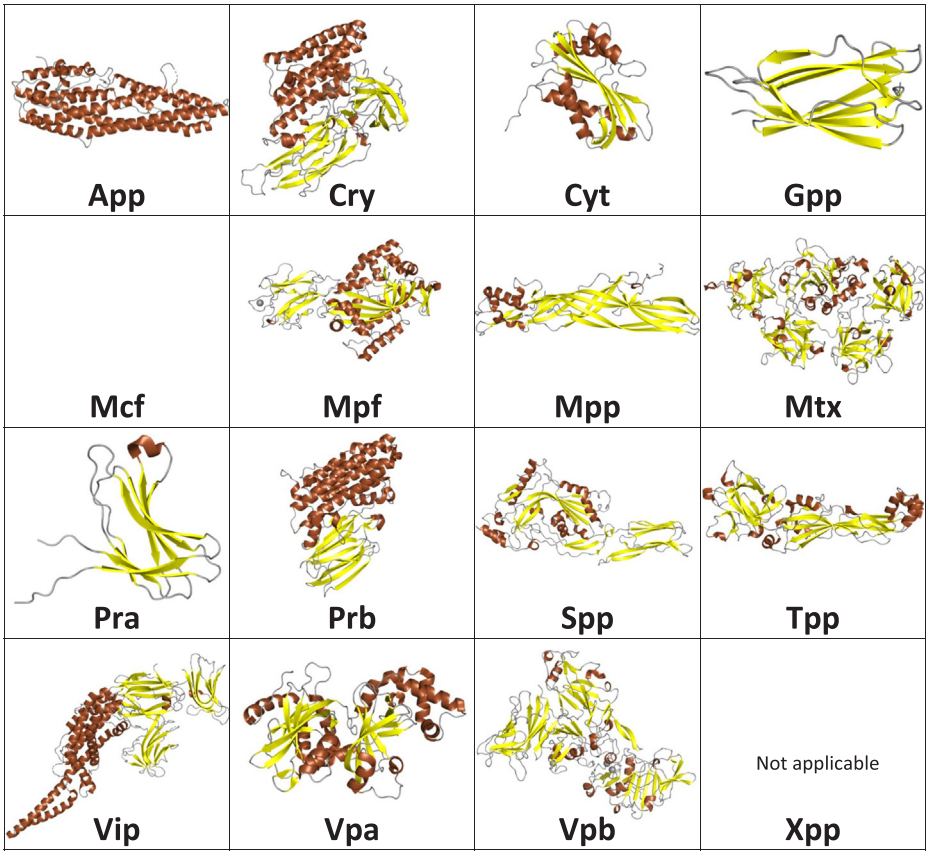

Fig. 1. Representative structures, where available, of the different pesticidal protein classes.

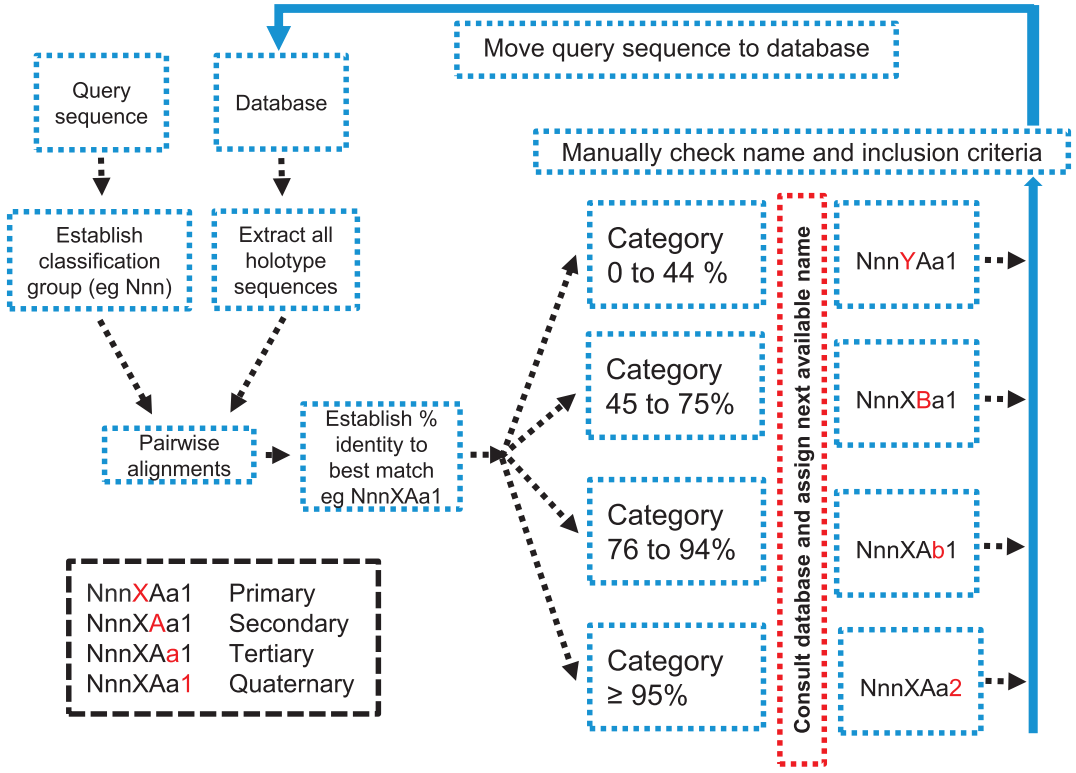

Fig. 2. Process used by nomenclature committee to name new bacterial pesticidal proteins.

Cry32Wa proteins have now been changed to Cry73Ba.

3.3. Use of Needle and adopting the best match approach for naming

Using the existing nomenclature as a scaffold, the simplest method for placing new sequences is to find the closest match protein and use the degree of identity to this protein sequence to derive the new name, and this proved to be as meaningful as any more complex method. For naming purposes, new sequences are only compared with holotype sequences within the nomenclature (i.e. those ending with the number 1). To perform the pairwise analyses, Needle (Madeira et al., 2019) is used in preference to BLAST (Altschul et al., 1990) as the former compares sequences over their full length rather than concentrating on shorter regions of homology in two diverse sequences. The classification level cut-offs used have remained very similar to those described in the 1998 protocol at 45%, 76% and 95% sequence identity. As with the 1998 nomenclature, full-length sequences are used for naming purposes rather than attempting to define functional regions. A flow diagram depicting the naming process is shown in Fig. 2.

3.4. Acquiring a new name for a pesticidal protein

Official pesticidal protein names will continue to be allocated by an appointed committee. The criteria for inclusion in the nomenclature, for proteins that would receive a holotype classification – i.e. ending with a 1, are that they should be derived initially from a bacterium, have demonstrated activity against a relevant pest species or target, and have had their coding sequence placed in a public repository (e.g. GenBank). We appreciate that researchers may not want to make sequences publicly available while publications or intellectual property are being prepared, in such circumstances both GenBank and the nomenclature databases can hold sequences securely until publication. For sequences that share 95% or more identity with a sequence already in the nomenclature, no demonstration of activity is required. Note that even though users can compare their own sequences against publicly available sequences in the nomenclature, official naming and addition to the nomenclature can only be undertaken by the committee.

4. Development of an interactive database and associated website

In association with the revised nomenclature, an online database has been set up, which can be accessed from www.bpprc.org. An interface to the database allows users to browse and download sequences as well as comparing their own sequences to those that are publicly available. As described above, users are able to request names for sequences that are not yet in the public domain. In such circumstances the names of these proteins will be listed but their sequences will not be available for viewing or searching. Where appropriate, these private sequences will be available to the nomenclature committee for naming purposes. In addition to users being able to search the database for the best matches to their own sequences, they will also be able to use the sequence comparison algorithm to compare two sequences, either from the database or supplied by the user. Other functionalities have been added including the ability to draw dendrograms of selected sequences from the database, with or without user sequences included. Such dendrograms can be derived from full length sequences or, for some protein classes, individual domains. The associated website also provides a portal for users to submit sequences for naming.

5. Concluding remarks

Although the increasing diversity of pesticidal proteins isolated from bacteria required a new look at the existing classification, there seemed little appetite for completely abolishing a system that has been widely accepted and adopted. This current revision has stuck to the original concept that the most important role of the nomenclature is to provide each protein with a unique identifier that can be used both in an academic context and through any commercialization activity. Although the given name does reflect its relatedness to other proteins within the nomenclature, it is not intended that the name specifically

indicates any particular functional or evolutionary characteristic. The current system has retained the original principle of giving each newly characterized sequence a unique identifier – even if the new sequence happens to be identical to an existing one. By attempting to minimise the changes to the previous nomenclature – while incorporating a new structure-based element – it is hoped that this revision will also be widely adopted. In this version, some new types of bacterial pesticidal protein have been incorporated into the nomenclature (see Table 1) and it is anticipated that new classes will continue to be added using the principles described above.

Funding

This material is based upon work supported by the National Science Foundation I/UCRC, the Center for Arthropod Management Technologies, under Grant Nos. IIP-1338775 and 1821914, and by industry partners.

References

Altschul, S.F., et al., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Crickmore, N., et al., 1998. Revision of the nomenclature for the *Bacillus thuringiensis* pesticidal crystal proteins. Microbiol. Mol. Biol. Rev. 62, 807–813.

de Maagd, R.A., et al., 2003. Structure, diversity, and evolution of protein toxins from spore-forming entomopathogenic bacteria. Annu. Rev. Genet. 37, 409–433.

Donovan, W.P., et al., 2006. Discovery and characterization of Sip1A: A novel secreted protein from *Bacillus thuringiensis* with activity against coleopteran larvae. Appl. Microbiol. Biotechnol. 72, 713–719.

El-Gebali, S., et al., 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47, D427–D432.

Estruch, J.J., et al., 1996. Vip3A, a novel *Bacillus thuringiensis* vegetative insecticidal protein with a wide spectrum of activities against lepidopteran insects. Proc. Natl. Acad. Sci. U. S. A. 93, 5389–5394.

Hofte, H., Whiteley, H.R., 1989. Insecticidal crystal proteins of *Bacillus thuringiensis*. Microbiol. Rev. 53, 242–255.

Hu, Y., et al., 2018. *Bacillus thuringiensis* Cry5B protein as a new pan-hookworm cure. Int. J. Parasitol. Drugs Drug Resist. 8, 287–294.

Liu, L., et al., 2019. Identification and evaluations of novel insecticidal proteins from plants of the class polypodiopsida for crop protection against key lepidopteran pests. Toxins (Basel) 11. https://doi.org/10.3390/toxins11070383.

Madeira, F., et al., 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 47, W636–W641.

Ohba, M., et al., 2009. Parasporin, a new anticancer protein group from *Bacillus thuringiensis*. Anticancer Res. 29, 427–433.

Schnepf, H.E., Whiteley, H.R., 1981. Cloning and expression of the *Bacillus thuringiensis* crystal protein gene in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. 78, 2893–2897.

Windley, M.J., et al., 2012. Spider-venom peptides as bioinsecticides. Toxins (Basel) 4, 191–227.